Computational Statistics
M.Sc. Mathematical Engineering
Prof. A. Manzoni

A.A. 2021 – 2022
Politecnico di Milano

Hands-On 1

Random Number Generation
Random Variable Generation - 1

Problem 1.1
A pseudo-random number generator is implemented in Matlab; the `rand` command allows us to generate real numbers (floating-point) uniformly distributed in the interval $[0, 1]$.

1. Let $X \sim \mathcal{U}([0, 1])$ be a uniform random variable in the interval $[0, 1]$. Let us sample X n times, with $n = 10^2, 10^3, 10^4, 10^5$. For each value of n:
   - calculate the sample mean and variance (using the `mean` and `var` commands) and compare them with the exact values of X;
   - display the histogram of the sampled values (`hist` and `bar` commands);
   - display the empirical distribution function of the data using the `cdfplot` command and compare it graphically with the cumulative distribution function of X.
2. Sample a random variable $Y \sim U([3, 23])$ and verify the correctness of the result by comparing the sample mean and variance with the theoretical values.
3. Let Z be a discrete random variable that can assume the integer values $1, 2, \ldots, 20$ with equal probability. Propose a method for sampling the variable Z.

Once `Matlab` is started, the first random number generated is always

```
>> rand
ans =
0.95012928514718
```

and the sequence of pseudo-random numbers will always be the same!

```
>> rand (1,3)
ans =
0.23113851357429 0.60684258354179 0.48598246870930
```

This is because the pseudo-random number generator depends on a state (consisting of 35 memory words accessible through the `S = rand ('state')` command) which is always initialized to the same value when `Matlab` is launched. At any time the generator can be restarted from the initial state by means of the command

```
>> rand ('state', 0)
>> rand (1,4)
ans =
0.95012928514718 0.23113851357429 0.60684258354179 0.48598246870930
```

This is extremely useful for verifying the correctness of a program (debugging). Conversely, if at each execution of a program you want to initialize the generator to one different state, you can use, for example, the following command:

```
>> rand ('state', 100 * sum (clock))
>> rand (1,4)
ans =
0.26053315553164 0.98934081890679 0.72014353997263 0.36942940565346
```

1. Recalling that, if $X \sim \mathcal{U}([a,b])$, then

$$\mathbb{E}[X] = \frac{a+b}{2}, \qquad \mathrm{Var}[X] = \frac{(b-a)^2}{12},$$

mean and variance of X are, respectively,

$$\mathbb{E}[X] = \frac{1}{2}, \qquad \mathrm{Var}[X] = \frac{1}{12} = 0.08\overline{3}.$$

Sample mean and sample variance are, on the other hand, implemented in `Matlab` in the `mean` and `var` commands

$$\mathtt{mean}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mathtt{var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mathtt{mean}(X))^2.$$

Let us consider the case $n = 100$ and run the following commands:

```
n = 100;
X = rand(1,n);
average = mean(X)
   average = 0.52856822744889

variance = var(X)
   variance = 0.07938991410138

figure(1);
dx=0.1; x=[0+dx/2:dx:1-dx/2];
N = hist(X,x);                     % hist conts how many elements of X fall in the
                                   % neighborhood of the points defined by x.
P = N/(n*dx);                      % approximation of the PDF
bar(x,P,1); hold on;
x = [0:0.01:1]; plot(x, 1+0*x, '-r');          % exact PDF
legend('Histogram','exact pdf',0)
hold off

figure(2);
cdfplot(X); hold on;                           % plot of the empirical CDF
x = [0:.01:1]; plot(x,x,'-r');                 % plot of the exact CDF
legend('empirical CDF','exact CDF',0)
hold off
```

hist counts how many elements fall into the partition

graph of the empirical cumulative distribution function

Figure 1 shows the histogram (on the left) and the empirical CDF (on the right).

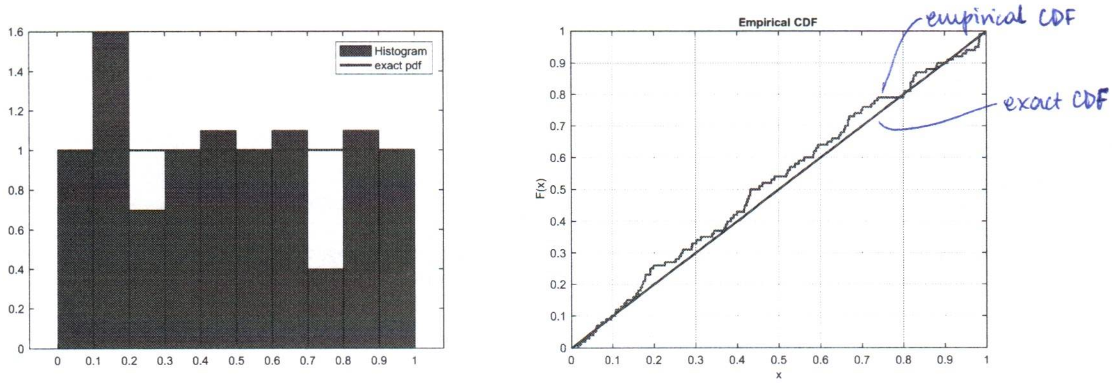

Figure 1: Histogram (left) and empirical CDF (right) for 100 values sampled from $\mathcal{U}([0,1])$.

Then, using the commands

```
test_cdf = makedist('Uniform','lower',0,'upper',1);
[h,p] = kstest(X,'alpha',0.05,'CDF',test_cdf)
```

we can test the null hypothesis that the sample has been extracted from a uniform distribution (defined through the command `makedist`);

MATLAB implementation of the Kolmogorov-Smirnov test

```
[h,p] = kstest(X,'alpha',0.05,'CDF',test_cdf)
```

returns a test decision for the null hypothesis that the data in vector X comes from the distribution `test_cdf`, against the alternative that it does not come from such a distribution, using the one-sample Kolmogorov-Smirnov test. The result `h` is 1 if the test rejects the null hypothesis at the `alpha` significance level, or 0 otherwise, and also the p-value `p` of the hypothesis test. In our case, we find `h=0` and `p = 0.2383`, hence we accept the null hypothesis.

However, from the graphs and from the values of sample mean and sample variance it can be seen that 100 values are not yet sufficient to have a representative sample of the variable X. Commands can be repeated for $n = 10^2, 10^3, 10^4, 10^5$.

2. The random variable Y can be obtained with a simple linear transformation of $X \sim \mathcal{U}([0,1])$: $Y = 3+20X$. In `Matlab`:

```
Y = 3+20*X;
mean(y)
    ans = 13.57136454897784

var(y)
    ans = 31.75596564055003
```

3. Let $W = 20X \sim ([0,20])$. The `ceil` command approximates a real number up to the nearest integer. It is immediately verified that the random variable `Z = ceil (W)` takes only integer values $1, 2, \ldots, 20$ with uniform distribution. For example, with 10^4 samples one finds

```
Z=ceil(20*rand(1,1e4));
figure(3); x=[1:20]; P=hist(Z,x)/n; bar(x,P);
figure(4); cdfplot(Z);
```

Figure 2 shows the histogram (on the left) and the empirical CDF (on the right).

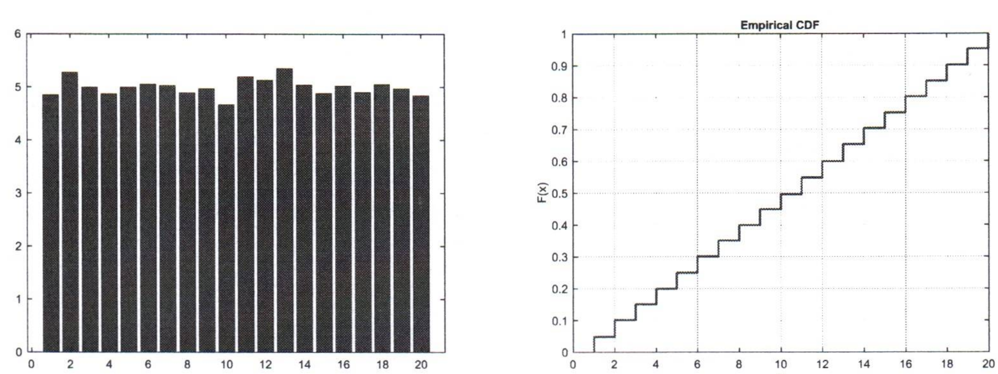

Figure 2: Histogram (left) and empirical CDF (right) for 10^4 values sampled from the discrete variable Z.

Problem 1.2 (Bernoulli)
We want to sample a Bernoulli random variable $X \sim Be(p)$:

$$P(X=1) = p, \qquad P(X=0) = 1-p.$$

1. Show that if $U \sim \mathcal{U}([0,1])$ is a uniform random variable in $[0,1]$, then X can be obtained as $X = \mathbf{1}_{\{U>1-p\}}$.
2. Draw 1000 samples of the variable $X \sim Be(p)$, with parameter $p = 0.3$, and check the correctness of the numerical results, comparing mean, variance and cumulative distribution function with the theoretical ones.

1. It is straightforward to show that

$$P(X=1) = P(U > 1-p) = p, \qquad P(X=0) = P(U \leq 1-p) = 1-p.$$

2. Mean and variance of X are given by

$$\mathbb{E}[X] = p = 0.3, \qquad \text{Var}[X] = p(1-p) = 0.21.$$

To draw a sample from X we can use the following commands:

```
n = 1000; p = 0.3;
U = rand(1,n);
X = U > 1 - p;
average = mean(X)
   average = 0.3190

variance = var(X)
   variance = 0.2175
```

without any other thing this returns n random numbers distributed iid $\sim \mathcal{U}([0,1])$

✓

Problem 1.3 (Inverse Transform method)
Let X be a continuous random variable taking values in $\mathbb{R}$ and $F : \mathbb{R} \to [0,1]$ the corresponding cumulative distribution function: $F(x) = P(X \leq x)$.

1. Show that the random variable $U = F(X)$ has a probability density which is uniform in $[0,1]$.
2. The previous point suggests a general way to sample X: sample a uniform variable $U \sim ([0,1])$ and set $X = F^{-1}(U)$. Use this technique to sample a continuous random variable X taking values in $[0,1]$, with probability density function $p(x) = 2x$. Check the correctness of the sample comparing the histogram and the empirical distribution function with the exact values.
3. Let Σ be the circle with center $(0,0)$ and radius 1. Let us denote by $\mathbf{x} = (x_1, x_2)$ a random point in the circle, with uniform distribution (that is, each point of the circle is equiprobable). Denoting by (r,θ) the polar coordinates, and setting $x_1 = r\cos\theta$, $x_2 = r\sin\theta$, which is the joint distribution of (r,θ)? Are r and θ independent from each other? Finally, take advantage of the transformation into polar coordinates to draw 1000 points int he unit circle, with uniform distribution.

1. We have

$$P(U \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y,$$

since F is monotonically increasing.

2. Let us denote by $f(x) = 2x$, $x \in [0,1]$, the PDF of X. The CDF is

$$F(x) = \int_0^x f(y)dy = x^2, \qquad 0 \leq x < 1;$$

$$F(x) = 0, \quad \text{for } x < 0; \quad F(x) = 1, \quad \text{for } x > 1.$$

The transformation $X = \sqrt{U}$ allows us to draw a sample from X from $U \sim ([0,1])$.

```
n=1000; Y=sqrt(rand(1,n));
dx=0.1; x=[dx/2:dx:1-dx/2];

figure(1); N=hist(Y,x); bar(x,N/(n*dx),1); hold on;
x=[0:0.01:1]; plot(x,2*x,'r'); hold off;

figure(2); cdfplot(Y); hold on;
plot(x,x.^2,'r--'); hold off;
```

Figure 3 shows the histogram (on the left) and the empirical CDF (on the right).

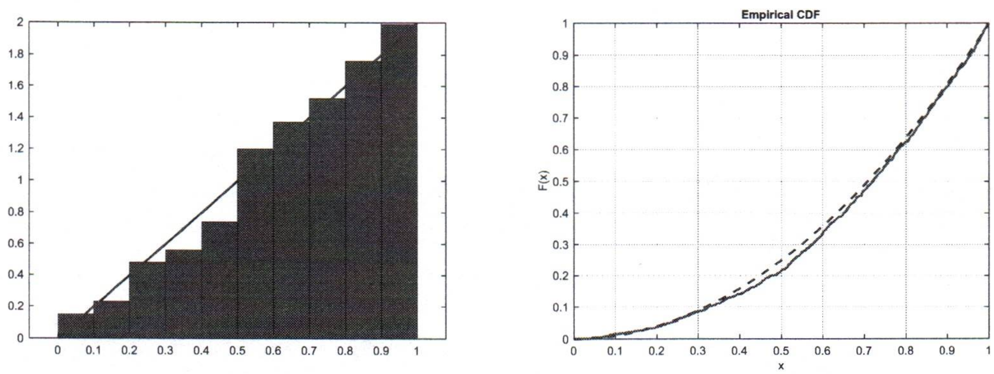

Figure 3: Histogram (left) and empirical CDF (right) for 10^3 values sampled from the variable X.

3. Since $\mathbf{x}$ is uniformly distributed on Σ, the PDF will be

$$f_{\mathbf{x}}(x_1, x_2) = \frac{1}{\pi} \qquad \forall (x_1, x_2) \in \Sigma$$

and using the transformation

$$G^{-1} : \begin{cases} x_1 = r\cos\theta \\ x_2 = r\sin\theta \end{cases} \qquad \text{with } r \in [0,1],\ \theta \in [0, 2\pi)$$

the joint PDF of (r, θ) will be (see Theorem 2.1)

$$f_{r,\theta}(r,\theta) = f_{\mathbf{x}}(x_1(r,\theta), x_2(r,\theta))|\det J| = \frac{1}{\pi} r = f_r(r) f_\theta(\theta)$$

where

$$f_r(r) = 2r, \qquad f_\theta(\theta) = \frac{1}{2\pi},$$

and having set

$$\det(J_{G^{-1}}) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

The joint distribution is therefore factored into the product of two distributions in the two variables r and θ, respectively. They are therefore independent and, as seen in the previous point, they are distributed as: $\theta \sim \mathcal{U}([0, 2\pi])$, and $r = \sqrt{u}$ with $u \sim \mathcal{U}([0,1])$. Points uniformly distributed in the unit circle can therefore be generated as follows (see Figure 4):

```
n=1000;
r=sqrt(rand(1,n));
t=2*pi*rand(1,n);
figure(3); plot(r.*cos(t),r.*sin(t),'o'); axis square; hold on;
fnplt(rsmak('circle'),'r'); hold off;
```

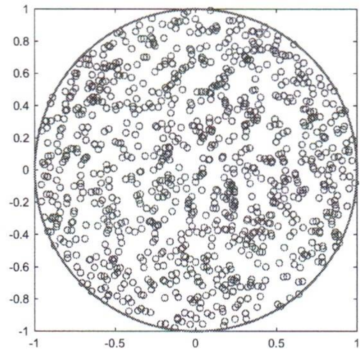

Figure 4: 1000 points uniformly distributed on the unit circle.

Problem 1.4 (Exponential and Gamma random variables)

1. Let $U \sim \mathcal{U}([0,1])$ be a uniform random variable in $[0,1]$. Show that the variable
$$X = -\frac{1}{\lambda}\log U$$
has an exponential distribution, $X \sim Exp(\lambda)$, with parameter λ.

2. Sample an exponential variable $X \sim Exp(2)$ and show that the sampling performed is correct.

3. We now want to sample the variable $Y \sim \text{Gamma}(m, \lambda)$, whose probability density function is
$$p(x) = \frac{1}{(m-1)!}\lambda e^{-\lambda x}(\lambda x)^{m-1}, \qquad z > 0, m \in \mathbb{N}.$$
Let us sample the variable $Y \sim \text{Gamma}(m, \lambda)$, with $m = 5$ and $\lambda = 2$, using the fact that Y can be obtained as the sum of m independent variables $Exp(\lambda)$. Show that the result is correct comparing the histogram of the sampled values with the probability density function defined above.

1. The PDF and the CDF of $X \sim Exp(\lambda)$, $\lambda > 0$, are
$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \qquad\qquad F(x) = \int_0^x \lambda e^{-\lambda y}\,dy = 1 - e^{-\lambda x},$$
respectively. The inverse of $y = F(x)$ is
$$1 - e^{-\lambda x} = y \quad \rightarrow \quad -\lambda x = \log(1-y) \quad \rightarrow \quad x = -\frac{1}{\lambda}\log(1-y)$$
and thus $F^{-1}(y) = -\frac{1}{\lambda}\log(1-y)$ for all $y \in (0,1)$. Now, assume $U \sim \mathcal{U}(0,1)$; we then obtain that
$$X = F^{-1}(U) = -\frac{1}{\lambda}\log(1-U)$$
is $Exp(\lambda)$-distributed. Moreover, since $1-U$ has the same distribution than U, an equivalent formula is
$$X = -\frac{1}{\lambda}\log(U).$$

2. With these commands

```
n = 1000; l = 2;
Y = -log(rand(1,n))/l;
dx= 0.1; x = [dx/2:dx:3-dx/2];
```

```
figure(1); N = hist(Y,x); bar(x,N/(n*dx),1); hold on;
x = [0:0.01:3]; plot(x,l*exp(-l*x), 'r'); hold off;

figure(2); cdfplot(Y); hold on;
plot(x,1 - exp(-l*x),'r--'); hold off;
```

we obtain the histogram (on the left) and the empirical CDF (on the right) in Figure 5

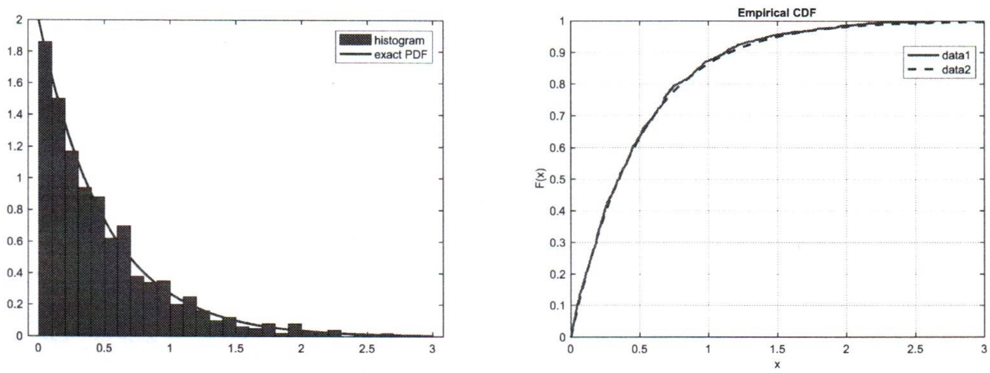

Figure 5: Histogram (left) and empirical CDF (right) for 10^3 values sampled from $X \sim Exp(2)$.

3. With the commands

```
n = 1000; m = 5; l = 2;
E = -log(rand(m,n))/l; % provides a matrix m x n of Exp variables
Y = sum(E);            % sum the matrix by column

dx= 0.4; x = [dx/2:dx:10-dx/2];
figure(3); N = hist(Y,x); bar(x,N/(n*dx),1); hold on;

x = [0:0.01:10]; p='l*exp(-l*x).*(l*x).^(m-1)/prod(1:m-1)';
plot(x,eval(p),'r'); hold off;

figure(4); cdfplot(Y); hold on;
plot(x,gamcdf(x,m,1/l),'r--'); hold off;
```

we obtain the histogram (on the left) and the empirical CDF (on the right) in Figure 6. In this case, the plot of the CDF has been obtained through the `gamcdf` command of the Statistical toolbox of `Matlab`.

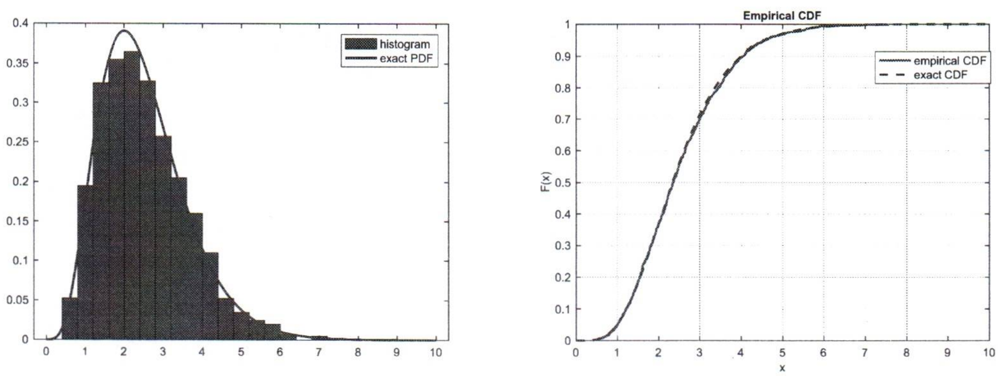

Figure 6: Histogram (left) and empirical CDF (right) for 10^3 values sampled from $X \sim Gamma(2, 5)$.